ELSEVIER

Cancer Detection and Prevention 27 (2003) 266–274

Cancer Detection and Prevention

www.elsevier.com/locate/cdp

Secular stability and reliability of measurements of the percentage of dense tissue on mammograms

Jacques Benichou, M.D., Ph.D. [a], Celia Byrne, Ph.D. [b], Laura A. Capece, B.A. [c], Leslie E. Carroll, B.A. [c], Kathy Hurt-Mullen, M.P.H. [d], David Y. Pee, M.Phil. [c], Martine Salane [e], Catherine Schairer, Ph.D. [f], Mitchell H. Gail, M.D., Ph.D. [f,*]

[a] *Department of Biostatistics, University of Rouen Medical School and Rouen University Hospital, 76031 Rouen Cedex, France*
[b] *Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA*
[c] *Information Management Services Inc., Rockville, MD 20852, USA*
[d] *Westat Corporation, Rockville, MD 20892, USA*
[e] *MSW Consulting, Bloomfield Hills, MI 48304, USA*
[f] *Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA*

Accepted 30 April 2003

Abstract

Elevated mammographic density is associated with increased risk of breast cancer. We conducted a reliability study on mammographic density assessments to determine their potential usefulness for projecting individual breast cancer risk. We used baseline screening mammograms from 7251 women in the Breast Cancer Detection Demonstration Project (BCDDP). Repeated measurements from the same images were used to assess measurement variability by an experienced evaluator. Intraclass correlations of assessments over time usually exceeded 0.9, indicating usefulness for prospective applications. Data also indicated it may be reasonable to include cases identified in the first year of screening together with other cases in developing a risk model. Older ages and increased weight were associated with decreased mammographic density. The density of the right breast slightly exceeded that of the left. Among women who developed breast cancer, the baseline mammographic density of the ipsilateral (diseased) breast was 0.53 (95% confidence interval (CI) 0.20–0.86) percentage units higher than in the contralateral breast.

Keywords: Breast cancer risk projection; Breast cancer risk factor; Measurement error; Assessing mammographic density

1. Introduction

Several studies have shown that the areal percentage of dense tissue on a mammograph, which we term mammographic density, is a strong predictor of breast cancer risk [1–4]. For example, Byrne et al. [3] found that the odds ratio for breast cancer exceeded four among women with mammographic density greater or equal to 75%, compared to women with a baseline mammographic density of 0%, which is comparable to the odds ratio conferred by having two first-degree affected relatives [5]. Thus, mammographic density may be among the strongest available non-invasive predictors of breast cancer risk [6]. Increased mammographic density is also more common than other strong risk factors for breast cancer and is therefore associated with a higher attributable risk. For example, Byrne et al. [3] estimated attributable risks of 28 and 46% for densities exceeding 49 and 0%, respectively.

Gail et al. [5] used data from the Breast Cancer Detection Demonstration Project [7] to develop a model for projecting the individualized absolute risk of breast cancer over a defined age interval based on number of affected first-degree relatives, age at first live birth, age at menarche, number of previous breast biopsies and the presence of atypical hyperplasia on a previous breast biopsy. This model, as modified by statisticians from the National Surgical Adjuvant Breast and Bowel Project [8], has been validated [8,9] and made available through the National Cancer Institute's Office of Cancer Communications and web site http://brca.nci.nih.gov/brc/. The present reliability and quality control study is part of an effort to determine whether the model can be improved by incorporating information on mammographic density, as preliminary data suggest [10].

Abbreviations: BCDDP, Breast Cancer Detection Demonstration Project; 95% CI, 95% confidence interval; S.E., standard error

* Corresponding author. Tel.: +1-301-496-4156; fax: +1-301-402-0081.
E-mail address: gailm@mail.nih.gov (M.H. Gail).

0361-090X/$30.00
doi:10.1016/S0361-090X(03)00092-8

This study uses mammographic density measurements from mammograms of 7251 women taken at the initial screening examination of the BCDDP between 1973 and 1975.

One precaution taken in some case–control studies to minimize the effects of temporal variability in assessing mammographic density is to measure mammographic density concurrently in matched cases and controls [3]. This strict control for temporal measurement variability is not possible when attempting to estimate a woman's risk prospectively. Thus, it is important to assess the degree of temporal variability in measurements, which can attenuate the strength of the association between mammographic density and subsequent risk. Case–control studies have relied on baseline images from the ipsilateral [3] or contralateral breast [4] to the breast with cancer in the case and used the corresponding image in the matched control. For risk prediction, one would not know laterality of disease. Thus, we evaluated the average baseline mammographic density from both breasts as a potential risk predictor in the present study.

The principal objectives of this study were: (1) to determine total measurement variability over time, including variability in outlining the dense area of breast tissue by an experienced evaluator and variability in subsequent processing by coders who use planimetry to measure the outlined area; (2) to identify and control for major factors that influence mammographic density when assessing variability in repeated measurements, and, in particular, age, weight, type of image (xeroradiogram versus film screen), and disease status; (3) to determine whether cases diagnosed in the first year beginning with the date of the initial screen in the BCDDP have comparable density to cases diagnosed in years 2–5 after the initial screen; and (4) to explore differences in density between the left and right breasts and between paired and unpaired assessments.

2. Methods

2.1. Study population

Between 1973 and 1975, 284,780 women entered the screening phase of the BCDDP and were followed for 5 years (up to 1980) with annual mammographic screening at 29 centers in the United States [5]. In a case–control study nested within the BCDDP cohort, cases diagnosed with invasive or in situ breast cancer were matched to controls on age at entry in 5-year intervals, race, study center, 6 month calendar time of screening, and length of follow-up [11]. In some instances, data were obtained on cases without matched controls and on controls without matched cases.

In 1980, a subcohort of 59,907 women who had not had a diagnosis of breast cancer during the screening phase was selected for further follow-up in phase I (1980–1986), phase II (1987–1989) and phase III (1993–1995) by telephone (phase I) and mailed questionnaires (phases II and III). The women in this subcohort included those with a biopsy diagnosis of benign breast disease during the screening phase (n = 25,114), those recommended for a surgical consultation but not biopsied (n = 9628), and a sample of "normal" women who had no breast surgery including biopsy or recommendation for surgical consultation during the screening phase (n = 25,165). These "normal" women were matched on age, date of entry into the BCDDP, race, center and length of participation to women with benign breast disease and breast cancer detected in the screening phase [12]. In a nested case–control study [3], incident cases within each of the three groups in the subcohort were matched with controls in the same group on follow-up time, center, race, and year of birth.

In all, 4275 women were found to have breast cancer during the screening phase, and 3090 during follow-up phases I–III. Only 28 centers provided information on breast cancer risk factors [3,7], and only 22 of the centers provided baseline mammographic images for the present study. The data in this paper are based on 2801 women in those 22 centers with incident breast cancer from the screening phase or from the subcohort follow-up phases I–III, together with 4450 controls.

Table 1 summarizes the numbers of women with mammographic density measurements from the baseline screening mammogram. In addition to the 6997 women with bilateral mammographic images, we used data from 254 women for whom only one breast image was available in some analyses (Table 1). Of all 7251 women, only the 7132 with weight

Table 1
Numbers of women with mammographic density measurements

	Women with both measurements (left and right breasts) available			Women with only one measurement available			All Women		
	Cases	Controls	Total	Cases	Controls	Total	Cases	Controls	Total
Screening phase (1973–1980)									
Cases detected in first year	505	503	1008	70	10	80	575	513	1088
Cases detected in second year	278	351	629	14	16	30	292	367	659
Cases detected in years 3–5	547	809	1356	24	26	50	571	835	1406
Phase I (1980–1986)	383	857	1240	16	19	35	399	876	1275
Phase II (1987–1989)	328	636	964	12	14	26	340	650	990
Phase III (1993–1995)	605	1195	1800	19	14	33	624	1209	1833
Total	2646	4351	6997	155	99	254	2801	4450	7251

measurements were used in regression analyses that included weight, age, case status and other factors. Among women with breast cancer, 2646 had baseline mammographic density measurements in both breasts, and an additional 155 had measurements on only one image (Table 1). Therefore, at least one mammographic density measurement was available for 38% of the 7365 breast cancer cases in the BCDDP study. Among the 575 cases diagnosed in the first year, 479 (83%) had evidence of breast cancer at the initial screening examination (prevalent cases).

2.2. Measurement of mammograms

The evaluator (M.S.), who was experienced in this technique, outlined the dense area of breast tissue on the cranio-caudal view with a wax-pencil [3]. We call this process "evaluating" or "assessing". Assessments of breast density were usually performed in pairs (left and right breasts), but some experiments investigated unpaired assessments. The evaluator was not informed of the case status or any other clinical data. Coders also masked to clinical status traced the breast perimeter and the dense area outline produced by the evaluator with a planimeter (Los Angeles Scientific Instrument Company, Model 1280-12) to estimate the area of dense tissue, the total breast area, and the ratio of the former to the latter times 100, namely the percent dense area. We call these planimetry measurements "coding". Variability in the measurement of percent density reflects variation in both assessment and coding. We calculated the average percent dense area of the two breasts. Unless otherwise noted, we use the term "mammographic density" or "density" to mean the average percent dense area from the two breasts. In those instances where only one breast image was available, its percent dense area was used as the mammographic density. In October 1999, the original planimeter was replaced with Model 1282W-12 from the Los Angeles Scientific Instrument Company. Mammograms were assessed from 15 September 1998 to 17 April 2000 in 88 batched evaluations, and the coding was accomplished from 13 November 1998 to 18 April 2000.

Images from 45 women were selected to represent the range of non-null breast density from the previous study [3] and included both film screen images and xeroradiograms. Coders used a set of 48 images from 24 of these women for initial training and the remaining 42 images from 21 women for testing the reliability of coding. These latter images, called the coder reliability sample, were used to assess coder performance and to quantify temporal variability attributable to coding. If a coder's performance was unacceptable, as defined in Appendix A, the coder was retrained, and if this proved ineffective, replaced. Coders were retested on the coder reliability sample at 6-month intervals. The outlined images in this sample were not cleaned between repeated readings. Thus, the same outlined markings were used repeatedly to assess temporal variability due to coding only. Coding variability also includes the effects of replacing the planimeter during the course of the study.

2.3. Reliability studies and other special studies

To evaluate reliability in repeated measurements over time (objective 1) as well as changes in density between the previous study [3] and the current study, 200 women were selected at random from among women previously studied by Byrne et al. [3]. We call this group of women the "overall reliability sample". Paired images from 197 of these women were assessed and measured at baseline in October and November of 1998 and 3 months later. The women in this sample were then randomly allocated into four subsets of approximately 50 women each. These subsets were assessed and measured in turn at six times separated by approximately 2-month intervals, so that two subsets were assessed and measured twice and two subsets were assessed and measured once. Assessments from the overall reliability sample and its subsets span the period 21 October 1998 to 6 March 2000. Unlike in the coder reliability sample, outlined images were cleaned between repeated readings and the evaluator had to outline dense areas again for each new reading.

To study factors that might influence density measurements (objective 2), we analyzed data on weight, age, case–control status, type of image and date of assessment from the 7132 women who had data on these factors and mammographic information from at least one breast image.

To determine whether the mammographic density of cases varied by time interval (in years) from the initial screening to the date of case diagnosis (objective 3), we studied cases diagnosed within first year of the initial screening ($n = 487$), in second year of screening ($n = 273$), and in years 3–5 of screening ($n = 543$). These numbers are smaller than the numbers in Table 1 because we also required complete information on weight, age, type of image and date of assessment. Here and throughout we use the phrases "within 1 year of the initial screening" or "year 1 of screening" to denote cases detected at the initial screening or within 1 year thereafter. Of the 487 cases detected in year 1 of screening, 82.5% had evidence of breast cancer at the initial screening.

To examine effects of laterality (objective 4), we computed correlations between the densities of left and right breasts and compared the mean densities of the left and right breasts in 4351 control women (Table 1) and in 2646 cases whose cancers were detected in the screening phase or phases I–III of follow-up. We also computed correlations between breasts and compared the densities of the diseased (ipsilateral) and non-diseased (contralateral) breasts in 1854 cases with unilateral disease. These cases excluded women diagnosed in phase III of follow-up, for whom available records do not preclude the possibility of bilateral disease, and some cases detected during the screening phase, phase I or phase II of follow-up, for whom the data on bilaterality was also unclear.

2.4. Statistical methods

Descriptive statistics and regression methods were employed. Linear regression models including fixed effects only or mixed linear models with fixed and random effects were used. The SAS General Linear Models procedure (PROC GLM) was used to study fixed effects regression models [13]. The SAS procedure for variance and intraclass correlation analysis (PROC MIXED) was used to estimate intraclass correlations while adjusting for fixed effects in mixed models [14]. Statistical significance was based on two-sided 0.05 level tests.

3. Results

3.1. Factors that influence mammographic density

We first examined factors hypothesized to affect mammographic density in order to identify those that would need to be controlled in subsequent reliability analyses. An analysis of results from 7132 women in the current study is summarized in Table 2. Cases had a density that was 6.87 percentage units higher on average than controls. Other factors, such as age and weight had even larger effects. The density in a 34–39-year-old woman exceeded that in a 70–77-year-old woman by 22.84 percentage units, and the density in a woman weighing 81–120 lbs exceeded that in a woman weighing 181–325 lbs by an estimated 31.16 percentage units.

Table 2
Parameter estimates (with standard error) from regression of mammographic density in percent on case status, age, weight, date of assessment and image type

Factor studied	Parameter estimate (standard error)[a]	Homogeneity test[b]
Case compared to control	6.87 (0.53)***	$P < 0.0001$
Age (compared to 70–77 years old)		
34–39	22.84 (1.82)***	$P < 0.0001$
40–49	20.39 (1.69)***	
50–59	10.68 (1.69)***	
60–69	5.66 (1.75)**	
Weight (compared to 181–325 lbs)		
81–120	31.16 (1.12)***	$P < 0.0001$
121–140	24.31 (1.04)***	
141–180	12.78 (1.05)***	
Date measured (compared to 1–60-day interval)		
61–120	0.57 (1.70)	$P < 0.0001$
121–180	−3.29 (1.82)	
181–240	−3.73 (1.75)*	
241–300	−2.66 (1.71)	
301–360	−1.78 (1.94)	
361–420	−0.33 (1.67)	
421–480	2.43 (1.69)	
481–540	5.03 (1.72)**	
Film screen vs. xeroradiogram	−0.94 (0.64)	$P = 0.139$
Intercept	−0.82 (2.45)	

[a] One asterisk denotes two-sided $P < 0.05$, two asterisks $P < 0.01$, and three asterisks $P < 0.001$. The analysis was based on 7132 subjects.

[b] Chi-square test of the homogeneity assumption that all the levels of this variable have the same effect on mammographic density. The degree-of-freedom equal the number of tabulated effects. For example, for weight, there are 3 d.f.

Secular variation in measurements of mean density was also evident. Compared to measurements in the period 1–60 days, measurements in the period 121–300 days were about 3 percentage units lower and in the period 421–480 days were about 2 percentage units higher. Those in the period 481–540 days were about 5 percentage units higher. Measurements in other periods were closer to the measurements in the initial period of 1–60 days. Some of this secular variation may be due to the type of women whose images were measured at various times in the study, even after adjusting for age, weight, image type and case status. However, some of the variation may be the result of changes in assessment technique or coding, as described in Section 3.2.

Homogeneity tests indicate statistically significant differences ($P < 0.0001$) among the levels of weight, age, date studied and disease status (Table 2).

Two types of images were used in this study, conventional X-ray film screens and xeroradiograms. Film screens yield mammographic densities 0.94 percentage units smaller than xeroradiograms, but this difference was not statistically significant.

3.2. Reliability studies

3.2.1. Temporal variation

To understand how much of the secular variation noted in Table 2 is attributable to assessment variability and coding variability, we analyzed mammographic density from women in the overall reliability sample, whose images were assessed multiple times over the course of the study (Fig. 1). The average change in mammographic density from the initial measurement at point A in November 1998 is indicated by the plus symbols in Fig. 1, whereas the solid circles give measurements for the ipsilateral breast only of the case or the corresponding breast for the control. By analyzing changes, we eliminate a source of variability in density measurements that arises from differences in mean densities among subgroups of the overall reliability sample. The average mammographic density changed by −1.81 percentage units at point B, 3 months after the initial measurement, and for the subsets of the overall reliability sample, the average change was −4.36 percentage units at point C (5 months), −5.13 percentage units at point D (7 months), 4.43 percentage units at point E (9 months), 0.07 percentage units at point F (12 months), −1.92 percentage units at point G (14 months) and 2.97 percentage units at point H (16 months). These data indicate a decrease in density from variations in the assessment and/or coding of the images 3–7 months following the initiation of the study and an increase at 9 and 16 months (Fig. 1). Similar changes were seen whether one

examined the ipsilateral breast only or the average mammographic density. Ipsilateral measurements made in 1994 using the original planimeter averaged about one percentage unit higher than the measurements at point A. Because the same women are being repeatedly analyzed when computing these mean differences, these changes are attributable to variations in the assessment and/or coding over time. We call variability due to assessment and/or coding the "total measurement variability". The reliability data on mammographic density in Fig. 1 suggest that most of the secular variation identified by regression analyses on all study participants in Table 2 is due to total measurement variability.

3.2.2. Reliability and intraclass correlation

To further analyze the total measurement variability, we studied three repeated measurements on each woman in the overall reliability sample, measured at time points A, B and one of the points C, D, E or F in Fig. 1. We estimated adjusted intraclass correlations by regressing the density measurements on all factors in Table 2 except "date read" and computing the intraclass correlations of the residuals from this regression. Adjusted intraclass correlations will tend to be smaller than unadjusted intraclass correlations, because the risk factors in the regression explain some of the between-woman variation in density. Using the data from this study (groups A–F), the correlations across time exceeded 0.90 in each case (left matrix in Table 3). Assuming a compound symmetric covariance matrix, namely a covariance matrix with equal correlations among all pairs of variates, we estimate the common correlation as 0.915 with 95% confidence interval 0.888–0.935. Very similar results on intraclass correlation were found in analyses of data from the ipsilateral breast only, at times A, B and C–F (right matrix in Table 3). Furthermore, the correlations between prior ipsilateral measurements, which were performed about 5 years earlier, and ipsilateral measurements at times A, B, and C–F were not much different (right matrix in Table 3).

From the coder reliability sample, we studied the residual correlation matrices after regression on image type (film screen versus xeroradiogram), case status, age (linear), weight (linear), and planimeter (old versus new). Under compound symmetry, the estimated correlations across five coding times at 6 month intervals were 0.999 for all three coders. The estimated residual variances were 0.47, 0.42 and 0.31 for the three coders respectively, about a 100 times lower than the 44.6 estimate for total measurement error obtained from the overall reliability sample. Because the evaluator's markings were unaltered from

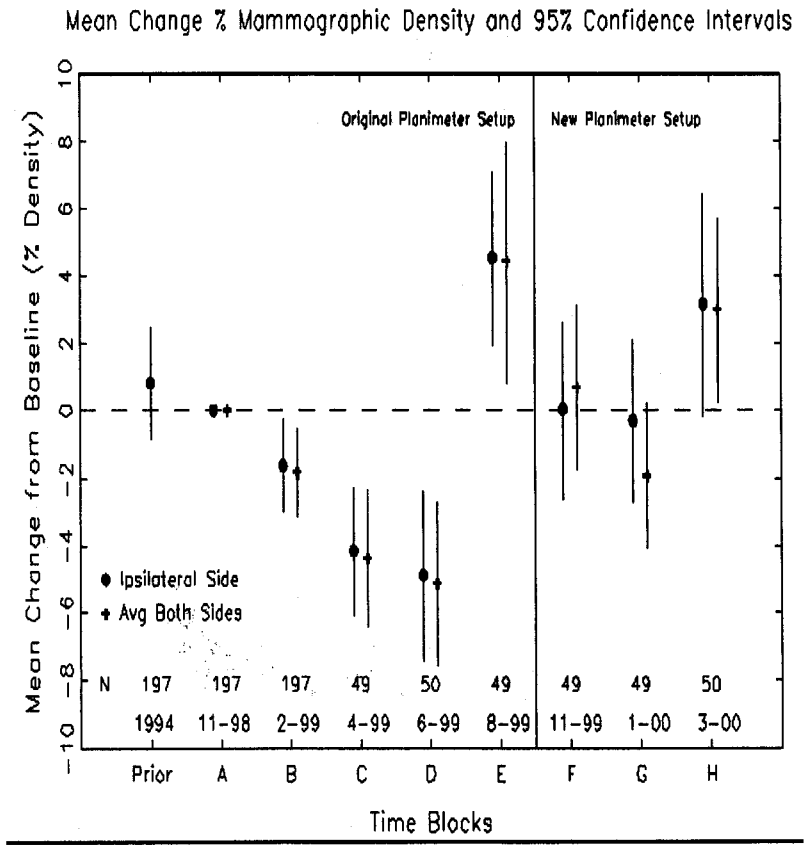

Fig. 1. Mean changes in mammographic density percentage units compared to baseline measurements at time point A for women in reliability sample 1 and, after time point B, for four subsets of these women examined in turn at times C–H. Vertical lines represent 95% confidence intervals, + signs indicate measurements based on the average of both breast images, and solid circles indicate results for the ipsilateral breast only.

Table 3
Residual covariances and correlations for mammographic density measurements among 193 women in the reliability sample with measurements separated by at least 2 months

	Both breasts			Ipsilateral breast only			
	A	B	C–F	Prior	A	B	C–F
Prior (1994)	–	–	–	586	514	468	467
A (11–98)	566	488	484	0.878	585	495	491
B (2–99)	0.915	503	468	0.858	0.908	507	466
C–F (4–99 to 11–99)	0.905	0.928	505	0.860	0.905	0.923	504

Diagonal terms correspond to variances of the residuals of the regression of mammographic density on age, weight, image type, and case–control status for measurements at various time points. Terms below the diagonal correspond to correlations of the residuals at different time points. Terms above the diagonal correspond to covariances of the residuals. These adjusted covariances are smaller than the unadjusted covariances and tend to yield smaller intraclass correlations than those based on the unadjusted covariances. The sample size is 193 because all covariates were required.

coding to coding in the coder reliability sample, these low values correspond to the variability attributable to planimetry and coding only. Thus, variation in planimetry and coding contributes extremely little to total measurement variability.

3.3. Possible sources of systematic error

3.3.1. Planimetry

Setting up and calibrating the planimetry coding system at the beginning of this study and switching planimeters during the course of this study had the potential to introduce systematic errors into the measurements. These possible biases appeared to be small, however (see Appendix B).

3.3.2. Time interval from initial screening to diagnosis

To determine if there were differences in the mean percent density by year of case diagnosis measured from the date of the initial screening, we regressed mammographic density on the factors in Table 2 among 1303 cases with required covariate data who were diagnosed in the first year of screening, the second year following the initial screening or in years 3–5 following the initial screening. The case indicator in the regression in Table 2 was replaced by indicator variables I_2 that took value 1 if the case was diagnosed in year 2 and 0 otherwise, and I_{3-5} that took value 1 if the case was diagnosed in years 3–5 and 0 otherwise. The corresponding estimates for effects of I_{3-5} and I_2 were 0.15 (S.E. 1.35) with $P = 0.91$ and 2.01 (S.E. 1.64) with $P = 0.22$, respectively. Moreover, the two degree-of-freedom test for any differences by year of diagnosis was not significant ($P = 0.42$). A similar regression analysis showed that the estimated density of 402 cases with evidence of breast cancer at the initial screening was 4.06 (95% CI −1.13, 9.25) percentage points lower than that of the 85 cases detected later in the first year, but the effect was not statistically significant ($P = 0.13$).

3.3.3. Laterality

The estimated correlation in baseline mammographic density between ipsilateral and contralateral breast images in 1854 women with unilateral breast cancer detected in the screening phase or phase I or phase II of follow-up was 0.957 (95% CI 0.953–0.961). The correlation between measurements from the right and left breasts was 0.960 (95% CI 0.958–0.963) in 4351 control women and 0.956 (95% CI 0.953–0.959) in 2646 women who developed unilateral or bilateral breast cancer in the screening phase or in phases I–III of follow-up. Interestingly, the percent area density of the right breast exceeded that of the left breast by 0.64 percentage units (95% CI 0.44–0.85) among the 4351 control women and by 0.72 percentage units (95% CI 0.44–1.00) among the 2646 cases.

We also examined whether the baseline mammographic density was greater in the ipsilateral (diseased) than in the contralateral breast of women with breast cancer. The mean ipsilateral mammographic density was 0.53 (95% CI 0.20–0.86 with $P = 0.0017$) percentage units higher than the contralateral mammographic density in 1854 unilateral cases. The unadjusted average ipsilateral mammographic density was 0.58 (95% CI 0.15–0.58) percentage units higher in 441 cases diagnosed in the first year, and 0.51 (95% CI 0.14–0.88) percentage units higher in 1413 cases diagnosed in years 2–5 of the screening phase and in follow-up phases I and II. These differences are small compared to the total measurement variability.

4. Discussion

This study was designed to assess whether mammographic density measurements obtained by an experienced evaluator were sufficiently reliable and free of systematic biases to be a useful risk factor for projecting individual breast cancer risk. Unlike earlier studies [3,4], we evaluated the average mammographic density of the two breast images, because the breast in which cancer may occur cannot be known when projecting breast cancer risk. In addition to studying the reliability (reproducibility) of measurements, we examined a number of potential sources of systematic error and factors thought to influence mammographic density. Age and weight have a large impact on mammographic

density (Table 2) and would need to be considered as potential confounders and modifiers of the risk associated with mammographic density in risk projection models. Mammographic density has been repeatedly shown to be inversely associated with age [2,3] and weight [2,3]. The effects of these and other factors on mammographic density and on mammographic patterns have been reviewed by Boyd et al. [15] and Byrne [16]. We obtained very similar results whether conventional film screens or xeroradiograms were used for imaging. The planimetry and coding of the traced dense areas contributed very little to total measurement variability. Different planimeter set-ups 5 years before the present study, at the beginning of the present study, and about 1 year later performed comparably.

A major concern was whether total measurement variability would attenuate the predictive value of mammographic density. In case–control studies, cases and matched controls can be measured at the same time, allowing one to minimize the impact of temporal variation on total measurement variability. This option is not available in prospective applications in which control for temporal variability is not possible. From data on the adjusted intraclass correlation, ρ, we estimated the degree to which total measurement variability attenuates estimates of risk in prospective studies based on a logistic model, as in Rosner et al. [17]. Let X represent the true mammographic density and X^* the measured mammographic density. Assuming we have stratified on the other covariates (except disease status and date read) in Table 2, one would anticipate that standard estimates of the log relative odds coefficient corresponding to X^* would be biased toward zero and would converge to $\beta^* = \rho\beta$, rather than to the value β that would be obtained if mammographic density were measured without measurement error [17]. Because the adjusted intraclass correlation is between 0 and 1, β^* is less than β in absolute value. Using data from an earlier study [3], we estimated β as 0.378 for a unit increase in percentage density category, with categories 0, 1–24%, 25–49%, 50–74%, and 75+% coded as 0, 1, 2, 3, and 4, respectively. In that study, breast density was assessed for the cases and their matched controls together. If we assume that there was no bias with regard to temporal measurement variability as a result, and if we take the estimate of ρ to be 0.915 as obtained from the compound symmetric model described Section 3, then $\beta^* = 0.915 \times 0.378 = 0.346$. Thus, the relative risk for the highest category compared to 0% density would be $\exp(4 \times 0.346) = 3.99$ instead of $\exp(4 \times 0.378) = 4.54$ in the absence of measurement error. This analysis suggests that mammographic density will remain a strong risk factor, even in prospective applications in which temporal variation contributes to total measurement variability.

Our intraclass correlation estimate was adjusted for case status, age, weight and film type, which would tend to produce a smaller value than unadjusted estimates. Nevertheless, our value of 0.915 appears consistent with unadjusted estimates given by Boyd et al. [4], who observed a correlation of 0.897, and Byrne et al. [18], who reported an unadjusted intraclass correlation of 0.93 for computer-assisted assessments.

The high intraclass correlations in our data indicated that most of the variabililty arises from biological differences among women, rather than from variation in assessments of mammographic density. In some applications, such as measuring the effects of hormone replacement therapy on mammographic density over time in the same women, it would be worthwhile to measure baseline and follow-up mammograms concurrently, to eliminate temporal variability in assessments, which can be appreciable (Fig. 1).

Egan and Mosteller [19] hypothesized that the association of density on screening mammography with subsequent cancer risk might be due, at least in part, to "masking" of tumors by dense tissue. Thus, prevalent cancers would be detected more readily at screening in women with low mammographic density but not in women with dense breasts at screening. In subsequent follow-up, those prevalent cancers not detected at the initial screen would be detected in the women with initially dense breasts, increasing the apparent incidence rate in such women. Our data indicate that mammographic density was slightly lower, but not statistically significantly so, for cases diagnosed within 1 year of the initial screening than for cases diagnosed in years 2–5 following the initial screening. Thus, cases diagnosed in year 1 may be useful for developing prospective risk models, and we plan to evaluate whether relative risks based on cases detected in the first year are consistent with those estimated from cases that arose later. Cases detected at the initial screening examination were about four percentage units less dense on average than cases detected later in the first year. Although this difference was not statistically significant, it is consistent with masking at the initial screening. The previous study in the BCDDP [3] excluded cases diagnosed at baseline and prior to the second screening exam. In that study, the relative risk for having $\geq$75% density was 7.58 (95% CI 3.2–17.9) for those diagnosed from 1 to 1.9 years after the baseline mammogram, but 4.47 (95% CI 2.1–9.6) for those diagnosed $\geq$10 years after the baseline mammogram. Although these variations are not statistically significant, they suggest that masking may play a role, but is unlikely to explain the entire association.

We note that the density measurements from the right breast exceeded those from the left breast by 0.64 percentage units (95% CI 0.44–0.85) in 4351 control women and by 0.72 percentage units (95% CI 0.44–1.00) in 2,646 women who developed breast cancer in the screening phase or phases I, II or III of follow-up. As suggested by a reviewer, the smaller percentage of dense tissue in the left breast may reflect its larger size [20], which determines the denominator in the calculation of mammographic density. The slightly higher density of the right breast that Byng et al. [21] and we observed was somewhat surprising, however, in view of incidence data from the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) program that demonstrates a 5% excess of left-sided breast cancers [22].

We obtained correlations above 0.95 in mammographic density between left and right breast sides both in cases and controls. To some extent, this agreement may reflect the fact that the evaluator assessed the paired breasts at the same time. Unfortunately, we were unable to determine from our data whether assessing images in pairs yielded different results from assessing unpaired images, because the unpaired data were confounded with temporal variability in measurements. Using unpaired images, however, Byng et al. [21] obtained similarly high correlations of 0.94 and 0.96 using a six-level visual assessment and slightly lower figures of 0.91 and 0.93 using computer-assisted and fully automated assessments respectively.

In women who developed breast cancer, the mammographic density of the ipsilateral breast exceeded that of the contralateral breast by 0.53 percentage units (95% CI 0.20–0.86). This difference suggests that a case–control analysis that relies on ipsilateral images only will tend to yield higher relative risks than an otherwise similar analysis that uses contralateral images only. The impact of this difference is not evident in comparing the results of Byrne et al. [3], who used ipsilateral images, and Boyd et al. [4], who used contralateral images, perhaps because of differences in adjustments for other factors in these two studies and perhaps because the laterality effect is small.

Several features of our study may limit the generalizability of the conclusions to other settings. First, this study was not designed to assess variability due to the mammographic image (i.e. variability between mammograms of the same breast at approximately the same date performed by different operators or at different X-ray facilities), since only one baseline mammogram was available for each woman. Second, our study relied on the assessments of a single highly trained evaluator (M.S.) to outline the dense areas on all mammograms. Measurement variability might have been greater if multiple evaluators had participated [15,16]. It is encouraging to note, however, that the current evaluator (M.S.) contributed only 22% of the readings in the previous study of Byrne et al. [3], to which two other evaluators also contributed, yet the intraclass correlations between the readings in the previous study and the current study were only slightly smaller than among various time points in the current study (Table 3). Third, we used the cranio-caudal view rather than the medio-lateral view to evaluate mammographic density. Using a computer-assisted technique, Byng et al. [21] demonstrated high correlations between measurements of breast density in cranio-caudal and medio-lateral views. Fourth, this study did not control for certain modulators of mammographic density, such as estrogen replacement therapy. Fifth, other measurement techniques, such as the computer-assisted technique described by Byng et al. [21], might yield different intraclass correlations from those we report.

This was a study of "internal" reliability to determine whether an experienced evaluator produced sufficiently reliable assessments of mammographic breast density over time for projecting individual breast cancer risk. This study does not address the comparability of the technique used to other methods of assessment, to evaluation by other experienced evaluators, nor to novice evaluators. Furthermore, our finding of high reliability does not imply that the assessments are unbiased with respect to some hypothetical absolute "gold" standard, which has yet to be defined.

Nonetheless, our findings indicate that assessment of the average mammographic density of the left and right breasts by an experienced evaluator are sufficiently reliable to serve as a useful predictor for projecting individual breast cancer risk.

Acknowledgements

We wish to thank Ms. Millie Bendell and her team of coders for their contributions to the study.

Appendix A. Evaluation of coders

Coders were tested initially and at 6-month intervals on the coder reliability sample of 42 images from 21 women. Coders traced the breast perimeter and the outlined dense area twice on each image in random order. The coefficient of variation was estimated for each coder on each mammogram. Acceptable performance for each coder required the mean coefficient of variation over the 42 mammograms to be 2% or less and the maximum coefficient of variation to be 6.54% or less. The figure 6.54% was obtained from simulations to determine the distribution of the maximum coefficient of variation when the true value was 2%. Acceptable coder performance also required a non-significant ($P > 0.05$) paired t-test comparing the first and second measurements from each mammogram. Also, the ratio of the maximum variance among the four coders to the sum of the variances of the four coders needed to be 0.49 or less, based on the 95th percentile of the simulated null distribution with equal variances. Otherwise, if acceptable performance were not obtained, the coder was retrained, and if this proved ineffective, replaced.

Appendix B. Effects of initially calibrating the planimetry system and switching planimeters during the course of the study

At the beginning of this study, the planimeter previously used in the study by Byrne et al. [3] was reset and calibrated. The average mammographic density in the ipsilateral breast in the reliability sample as measured in the study of Byrne et al. [3] was 36.45, which is only slightly larger than the value 35.64 found at time point A in the current study (Fig. 1). The difference in mean values, 0.81 had a standard error (S.E.) of 0.85, and the corresponding one-sample *t*-test

with $n - 1 = 197 - 1 = 196$ d.f. yielded $P = 0.34$. Thus, there was no statistically significant evidence that reassembling and calibrating the planimeter system after an interval of about 5 years affected the results.

Because the planimeter that had been used in the previous study and at the beginning of the current study became difficult to control beginning in September, 1999, a new planimeter was installed in October, 1999. As another test of the impact of a change in planimeter, we considered data from the coder reliability sample of 21 women whose outlined images were repeatedly coded throughout the study, and we compared the average codings before and after the planimeter change. The mean differences (new − old planimeter) were −0.11 with S.E. = 0.11 and $P = 0.33$ for coder 1, −0.53 (S.E. 0.18, $P = 0.007$) for coder 2, and 0.14 (S.E. 0.13, $P = 0.28$) for coder 3. Thus, there was no consistent evidence that the change in planimeter introduced an appreciable bias.

References

[1] Wolfe JN, Saftlas A, Salane M. Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case–control study. Am J Roentgenol 1987;148:1087–92.
[2] Saftlas AF, Hoover RN, Brinton LA, et al. Mammographic densities and risk of breast cancer. Cancer 1991;76:2833–8.
[3] Byrne C, Schairer C, Wolfe J, et al. Mammographic features and breast cancer risk: effects with time, age, and menopause status. J Natl Cancer Inst 1995;87:1622–9.
[4] Boyd NF, Byng JW, Jong RA, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst 1995;87:670–5.
[5] Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–86.
[6] Byrne C. Studying mammographic density: implications for understanding breast cancer. J Natl Cancer Inst 1997;89:531–3.
[7] Baker LH. Breast cancer detection demonstration project: five-year summary report. CA 1982;32:194–225.
[8] Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541–8.
[9] Rockhill B, Spiegelman D, Byrne C, et al. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001;93:358–66.
[10] Benichou J, Byrne C, Gail MH. An approach to estimating exposure-specific rates of breast cancer from a two-stage case–control study within a cohort. Stat Med 1997;16:133–51.
[11] Brinton LA, Hoover RN, Fraumeni Jr JF. Menopausal oestrogens and breast cancer risk: an expanded case–control study. Br J Cancer 1986;54:825–32.
[12] Velie E, Kulldorff M, Schairer C, et al. Dietary fat, fat subtypes, and breast cancer in postmenopausal women: a prospective cohort study. J Natl Cancer Inst 2000;92:833–9.
[13] SAS/STAT User's guide, version 6, vol. 2. Cary (NC): SAS Institute; 1990.
[14] SAS Technical Report P-229. SAS/STAT software: changes and enhancements, release 6.07. Cary (NC): SAS Institute; 1992.
[15] Boyd NF, Lockwood GA, Byng JW, et al. Mammographic densities and breast cancer risk. Cancer Epidemiol Biomarkers Prev 1998;7:1133–44.
[16] Byrne C. Mammographic density and breast cancer risk: the evolution of assessment techniques and implications for understanding breast cancer. Semin Breast Cancer 1999;2:301–14.
[17] Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051–70.
[18] Byrne C, Colditz GA, Willett WC, et al. Plasma insulin-like growth factor (IGF) I, IGF-binding protein-3, and mammographic density. Cancer Res 2000;60:3744–8.
[19] Egan RL, Mosteller RC. Breast cancer mammography patterns. Cancer 1977;40:2087–90.
[20] Senie RT, Saftlas AF, Brinton LA, Hoover RN. Is breast size a predictor of breast cancer risk or the laterality of the tumor? Cancer Causes Control 1993;4:203–8.
[21] Byng JW, Boyd NF, Little L, et al. Symmetry of projection in the quantitative analysis of mammographic images. Eur J Cancer Prev 1996;5:319–27.
[22] Weiss HA, Devesa SS, Brinton LA. Laterality of breast cancer in the United States. Cancer Causes Control 1996;7:539–43.